Lung Cancer Diagnosis Extraction from Clinical Notes Written in Spanish

1st Oswaldo Solarte-Pabon
Centro de Tecnología Biomédica
Universidad Politécnica de Madrid
Madrid, Spain
oswaldo.solartep@alumnos.upm.es

2nd Maria Torrente
Hospital Universitario Puerta de Hierro
Madrid, Spain
maria.torrente@salud.madrid.org

3rd Alejandro Rodríguez-Gonzalez
Centro de Tecnología Biomédica
Universidad Politécnica de Madrid
Madrid, Spain
alejandro.rg@upm.es

4nd Mariano Provencio
Hospital Universitario Puerta de Hierro
Madrid, Spain
mariano.provencio@salud.madrid.org

5nd Ernestina Menasalvas
Centro de Tecnología Biomédica
Universidad Politécnica de Madrid
Madrid, Spain
ernestina.menasalvas@upm.es

6ndJuan Manuel Tuñas
Centro de Tecnología Biomédica
Universidad Politécnica de Madrid
Madrid, Spain
juan.tunas@ctb.upm.es

***Abstract*—The wide adoption of electronic health records (EHRs) provides a potential source to support clinical research. Lung cancer is one of the most common cancers in the world. Although several tools have been developed to automatically extract medical concepts from clinical notes, there is still a gap between concept extraction and concept understanding. The high number of clinical notes written for one single patient, the use of negation, speculation and proper date annotations, lay at the root of the problem. In this paper, we propose an approach to obtain more accurate Lung cancer diagnosis extraction from clinical notes written in Spanish. The approach deals with a disambiguation process that is required to extract the correct date and diagnosis of a patient having hundreds of clinical notes and, consequently hundreds of annotations. Results obtained on an annotated database of 1000 patients show an F-score of 84%.**

***Index Terms*—Natural Language Processing (NLP), Information extraction, Lung cancer Diagnosis, Diagnosis extraction.**

I. Introduction

Lung cancer is one of the most common chronic diseases in the world and the leading cause of cancer death among both, men and women [1] [2] [3]. Accurate identification of lung cancer related information is crucial to support clinical and epidemiological studies, especially in terms of prognosis [4]. Giving greater attention to cancer diagnosis is a key factor for both the effective control of the disease, as well as the design of treatment plans. [5].

The increase of EHRs in recent decades opens up the possibility of analyzing them to extract information hidden in clinical notes [6]. These notes contain potentially useful and valuable information to support clinical decision making [7] [8]. However, the information in these EHRs is in a free-text form, which makes the task of data structuring them challenging. Despite the efforts of Natural Language Processing (NLP) tools to annotate medical concepts, once clinical notes are annotated, an additional process is required to integrate a patient's information over time. The process has to deal with a huge number of clinical notes written by different professionals in which repetition is common.

In SNOMED [9], one can find CUIs (concept identifiers) for Lung cancer, but not all of them are equally precise. This is due to the fact that the final diagnosis in which cancer stage annotation is found, only comes after pathology results are delivered. Meanwhile, mentions to the diagnosis can be found, but they may not be precise. Additionally, mentions to dates do not necessarily mean having the exact date implicit in the text, as it is common to refer to previous pathology, tests, or periods of a patient's natural history. Lastly, mentions to diagnosis can also appear as suspicion, or even negative after a test or procedure is executed.

In this paper, we propose an approach to establish the most accurate annotation for lung cancer diagnosis, together with the date of diagnosis. The approach starts from a traditional NLP process, that has been run in a system based on Apache UIMA [10]. We then enhance the NLP process with an additional step that disambiguates extracted annotations. The proposed approach takes advantage of the following annotators: i) diagnosis, ii) TNM[1], iii) negation and speculation, and iv) dates. Besides, the approach takes into account the kind of clinical note (family history, diagnosis, treatments, . . .). Several heuristics have been applied and obtained results show that the application of the negation and speculation annotators improve the performance to extract the cancer diagnosis and the diagnosis date.

The rest of the paper has been organized as follows: Section II reviews the most recent works on Lung cancer concepts annotation, Section III reviews the main challenges for dealing with accurate Lung cancer diagnosis. Section IV describes the components of the proposed approach. Section V presents the

This paper is supported by European Union's Horizon 2020 research and innovation program under grant agreement No. 727658, project IASIS.

[1] https://www.cancer.gov/about-cancer/diagnosis-staging/staging

2372-9198/20/$31.00 ©2020 IEEE
DOI 10.1109/CBMS49503.2020.00099

results of the experiments, and Section VI presents the main conclusions and outlook for future work.

II. Related Works

The use of NLP techniques to extract information related to cancer from clinical notes, has grown in recent years because it can be used as a tool for oncology evidence-based research and quality improvement [11].

One of the first interests was to extract Cancer stage, one important prognostic factor to understand cancer-specific survival. In [12] an NLP algorithm was developed to extract stage levels (I, II, III, IV) from clinical documents, including automated rules to choose the most likely stage when there was ambiguity in the EHR.

In [13] it is proposed an automated extraction tool for TNM cancer staging from free text pathological reports of breast cancer patients. TNM is extracted from pathological reports from different hospitals using a combination of pattern matching and rule-based techniques. Other proposals relating to studies to extract and structure cancer staging data are described in [14] and [15].

Recently other proposals have attempted to extract more concepts associated with the diagnosis of cancer and map it to concepts in the Unified Medical Language System (UMLS) Metathesaurus [16]. This mapping aims to standardize the extracted information. In [17] the authors propose a strategy to extract breast cancer diagnosis, History of malignant neoplasm and, if it is a recurrent cancer event.

On the other hand, [4] describes a system to extract information about cancer stage, histology, tumor grade, and therapies (chemotherapy, radiotherapy, surgery). Clinical notes, pathology reports, and surgery reports are used to test the system. The authors use time windows of 30, 60 and 90 days to measure precision and recall for histology and tumor grade of lung cancer diagnosis.The authors highlight the feasibility of extracting cancer-related information from narrative EHR data and the feasibility of improving the efficiency of humans through NLP techniques.

Although the proposals mentioned above showed significant advances extracting cancer-related information, and several systems was been developed [18] [19], most of these approaches have focused on the English language. According to [20], information extraction in the medical domain also represents its own challenges in languages other than English.

In the Spanish language, [5] proposes a system of automatically extracting concepts such as stage, performance status and mutations in the Lung cancer domain.

To deal with the recognition of time expressions, in [21] a temporary Tagger Annotator is proposed. This tool aims to identify and normalize time expressions in Spanish clinical texts. However, according to [22], extracting concepts and temporal expressions is not sufficient to understand events relating to patients. This is because extracted concepts can contain ambiguities, and require an additional process to bridge the gap between concept extraction and concept understanding.

III. Challenges for Annotating Cancer Diagnosis

When a patient goes to a hospital with suspected lung cancer, different tests are required to confirm the final diagnosis. During all the process of the patient many interactions occur of the doctor and the patient and consequently different notes are generated in which the doctor will report on the patient antecedents (some of which can be cancer), physical status and the suspected diagnose. When finally a patient is diagnosed this will be written in a note or report and it can occur that a explicit mention to the date will be written or simply the diagnose is written and the date is implicit to the date of the note. From the moment of the diagnose on, the physician in his interaction with the patient will write multiple notes and in most of them will make reference to the diagnosis, the way in which it will write the diagnose can vary (lung cancer, carcinoma, ca, ...) and so will be the way in which the date of the diagnose from implicit to relative dates to other moments of the natural history of the patient(24/03/98, three days before treatment, may this year, ...).

Although NLP tools and in particular entity recognition process facilitate the automatic annotation of clinical narratives, a post-process will be required first of all to eliminate those annotations that are either suspicion or negated and on the other those ones that do not refer to the patient himself (familial antecedents). To end with, in order to find the accurate date of diagnose all the diagnose references will be analyzed and heuristics will be applied in the case of multiple annotations to find the most appropriate one.

A. Annotations with different dates

Identifying the correct diagnosis date is a challenge since many date annotations can be extracted. Setting the correctly diagnostic date is affected by the causes mentioned above. Therefore, diagnoses that refer to the patient must be selected first and then the mechanism to choose the date will be established. The following example demonstrates this:

- *"On January 12 2017, LSI tests were recommended, suspicious Lung carcinoma."*
- *"Father operated of Lung cancer in September 2014.*
- *Patient with Small Cell Lung carcinoma in May 2017*

In all this cases we will find after the annotation process sentences in which we can find a date and a diagnose. However only in the last case the diagnose refers to the patient.

IV. Approach for extracting cancer diagnosis

This approach aims to extract the lung cancer diagnosis and diagnosis date from clinical notes. The approach consists of three steps: NLP annotation, Disambiguation process and, Diagnosis extraction. Figure 1 shows the proposed approach.

A. Step 1: NLP Annotation

This step refers to the NLP process in which the dates and the diagnosis are annotated. This process also annotates other medical concepts but we focus on these ones in this paper. In particular the approach takes into account all the clinical notes of a patient and clusters them according to their type. Once

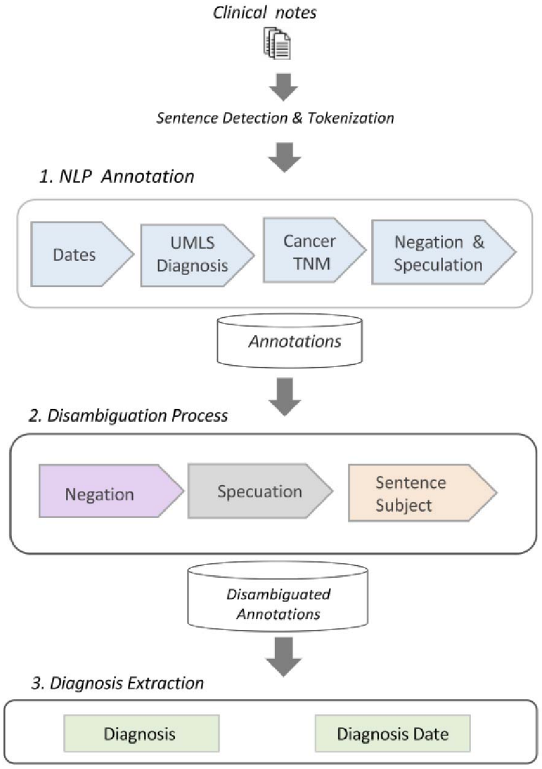

Fig. 1. Diagnosis Extraction Approach

the clinical notes regarding to clinical judgement are found they are chronologically ordered.

This step takes as input clinical notes that were previously divided by sentences and tokenized. As output, it generates a database with cancer-related annotations. The annotation process is performed using C-LIKES, a clinical information extraction system developed to process documents written in Spanish [23]. The following annotators are taken into account:

- **Dates Annotator:** This annotator recognizes dates and time expressions and transforms them into a standardized way. Dates annotator is able to structure expressions about dates written in different formats and styles in the Spanish language, as described in [24].
- **Diagnosis Annotator:** This annotator uses UMLS Metathesaurus [16] to recognize concepts associated with a cancer diagnosis. The diagnosis is annotated using UMLS, therefore the Concept Unique Identifier (CUI) is obtained for each diagnosis found in a clinical note. Each UMLS annotation links cancer diagnosis concepts with a CUI in the Metathesaurus.
- **TNM Annotator**: Extracts cancer stage using the Classification of Malignant Tumors or TNM notation. This notation represents the stage using three alphanumeric codes: Tumor(T), describes the size of the tumor. Nearby, describes lymph nodes (N) that are involved. Metastasis (M) describes the spread of cancer from one part of the body to another. (e.g *Patient with Lung Cancer TNM: cT3cN3cM1*). Extracting the cancer stage is relevant because it is an important indicator for determining the cancer diagnosis.
- **Negation & Speculation Annotator:** This annotator identifies negated and speculated medical concepts. As an input, it takes a previously annotated medical concept and generates another annotation that indicates if the concept is negated or speculative.
  A negation annotation happens when physicians exclude diagnoses by negating them. A speculative annotation occurs when the diagnosis is uncertain or doubtful.
  Negation annotator is based on the approach proposed by [25]. This approach is focused on the way in which negation is expressed in clinical notes written in Spanish.

Although NLP Annotation automatically extracts cancer diagnosis concepts, this step can generate ambiguous information. Table 1 shows a set of sentences and extracted concepts using NLP-based annotators. These sentences generate ambiguous annotations as a consequence of the following facts:

- There are many ways of indicating Cancer:*"Cancer, Carcinoma, Adenocarcinoma, Neoplasm, Ca, . . ."*
- Different dates are annotated: This increases ambiguity to extract diagnosis date, an important factor to understand the patient's timeline.
- Some sentences do not contain dates annotations or contain incomplete dates specification. (e.g *2016/01*).
- Sentences contain words indicating negation or speculation.(e.g *" (Rule out, Discard, Suspect, Suggest, . . . ")*
- Extracted concepts from the text may refer to other subjects and not to the patient. *(e.g "Brother, Father, Mother, . . .")*
- Many annotations relating to a cancer diagnosis can be extracted for each patient. Figure 2 shows a distribution of the number of annotations obtained for a sample of 1000 patients. This can increase the ambiguity about choosing the exact diagnosis date for the patient.

The facts shown above indicate that annotations obtained through NLP-based annotators contain ambiguous information. That is why a disambiguation process is needed to extract correct values for cancer diagnosis.

B. Step 2: Disambiguation Process

The main goal of this step is to disambiguate the multiple annotations that are due to the fact that, for the same patient multiple notes are written.

This process uses three components to disambiguate annotations obtained previously:

- **Negation Disambiguation:** This component filters annotations with negated diagnosis concepts. (e.g *"Negative liquid biopsy of lung Adenocarcinoma" ("Biosia líquida negativa para Adenocarcinoma de pulmón."*)

TABLE I
SENTENCES AND ANNOTATED DIAGNOSIS CONCEPTS

Date	Concept	Sentence
2016-07	Lung Cancer	Treatment: Lung Cancer in (July-2016).
2016-01	Lung Neoplasm .	**Suspicious** Lung Neoplasm 2016/01.
	Lung Adenocarcinoma.	62 year old woman with Lung Adenocarcinoma.
	Adenocarcinoma	Biopsy test **compatible** with adenocarcinoma
2012-3	Lung cancer	**Father** with Lung cancer in March 2012
	Lung Adenocarcinoma	Lung Adenocarcinoma Stage III
2016-05-01	Lung adenocarcinoma	Lung adenocarcinoma **cT3 cN3 cM1** (2016/05/1)
	Pulmonary neoplasia	Findings **suggest** primary pulmonary neoplasia
2017-2-17	Lung Cancer	Lung Cancer Follow-up (2017-2-17)
2015-09-10	Lung cancer.	Test to **rule out possible** lung Cancer.

- **Speculation Disambiguation:** Filters annotations with speculated diagnosis concepts. (e.g *"Suspicion of Lung neoplasm, 2016-01-12" ("Sospechosa neoplasia pulmonar, 2016-01-12")*
- **Sentence subject Disambiguation:** Filters annotations that do not belong to the patient as a subject. That is annotations that mention family history(e.g *"His father suffered Lung cancer in march 2012)." ("Padre con cáncer de pulmón en Marzo de 2012)*

Although the above examples contains diagnosis concepts *(Lung Adenocarcinoma, Lung neoplasm, Lung cancer)*, these examples do not contain a correct diagnosis date because they are negated or speculated or does not belong to the patient. The disambiguation process automatically generates a new database with disambiguated annotations. This database contains diagnosis annotations without negation, without speculation nor family history. Disambiguated annotations will be used in the next step to extract cancer diagnosis.

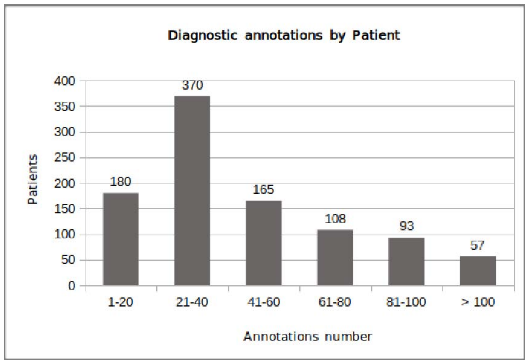

Fig. 2. Number of diagnosis annotations for each patient

C. Step 3: Diagnosis Extraction

In this step, the cancer diagnosis and the diagnosis date are extracted from disambiguated annotations obtained in the previous step.

- **Extracting diagnosis Date:** To extract the diagnosis date correctly, first clinical notes are classified according to their clinical group: (*Clinical Judgment, Medical Evolution, Surgery, Medical emergency, etc*).
  Clinical Judgement notes (**CJ**) are prioritized in this step. Figure 3 shows a rule-based algorithm to extract the diagnosis date. This algorithm is based on the next heuristics:
  - *Clinical Judgment notes (CJ)* commonly contain more diagnosis annotations than other clinical notes. For this reason CJ are prioritized to extract cancer diagnosis.
  - *TNM concept* is commonly used when physicians give to the patient a diagnosis. In this case, the first date mentioned in TNM annotations is chosen.
  - If the patient does not contains annotations in Clinical Judgment notes, the diagnosis date is extracted from other clinical notes.
- **Extracting diagnosis name:** To extract the diagnosis name our approach uses a ranked list of UMLS CUIs. This list contains UMLS codes ordered according to those that describe more completely the diagnosis name. Table 2 shows five diagnosis concepts and their respective CUI. These concepts are ordered from highest to lowest according to their relevance. Therefore, those annotations that coincide with the concept in the first row, will be more relevant.

V. EXPERIMENTS

A. Dataset and tests

To validate our approach, we use a database containing data from *"Hospital Universitario Puerta de Hierro Madrid"*

```
For each Patient do {
    Find TNM annotations in CJ

        If TNM annotations are found
            Chronologically order them
            Choose the first mention

        If Not TNM annotation are found
            Find diagnosis annotations in CJ
            Chronologically order them
            Choose the first mention

        Else
            Find diagnosis annotations
            Chronologically order them
            Choose the first mention
}
```

Fig. 3. Algorithm to diagnosis date extraction

TABLE II
A RANKED LIST OF DIAGNOSIS NAMES

CUI	Concept
C0001418	*“Small cell Lung Adenocarcinoma”*
C0006826	*“Lung Cancer”*
C0027651	*“Lung Neoplasm”*
C0027651	*”Cancer”*
C0006826	*”Ca”*

(HUPHM), in the last 10 years. It contains around 300,000 clinical notes corresponding to 1000 patients diagnosed with lung cancer. HUPHM hospital also contains a manually annotated dataset with the lung cancer diagnosis and their date for each patient. This dataset annotated by human experts is used for validating our experiments.

In the following, we describe the tests performed to calculate the performance of the proposed approach. We will analyze the impact on the accuracy of adding each disambiguation components to extract the diagnosis date. The algorithm shown in Figure 3 is used to extract the diagnosis date.

- **Test1**: Analyzes the accuracy of using all annotations generated by the NLP-based annotators. Neither disambiguation component is used, annotations are not filtered.
- **Test2**:Negation component is added to filter negated diagnosis annotations.
- **Test3**: Speculation component is added to filter annotations with suspicion concepts.
- **Test4**: Sentence subject disambiguation is added to filter annotations about family antecedents. Note that in Test 4, the algorithm uses annotations filtered by all disambiguation components: Negation, Speculation, and Sentence subject.

B. Metrics

The main goal of the tests is to measure the performance of the approach when extracting diagnosis dates. We consider that a diagnosis date is correct when the approach extracts this date exactly as established by the physician in a clinical note. In order to measure the performance we will use Precision, Recall, and F-score as follows:

$$\textbf{Precision} = \frac{\text{Diagnosis dates correctly extracted}}{\text{Diagnosis dates extracted}} \qquad (1)$$

$$\textbf{Recall} = \frac{\text{Diagnosis dates correctly extracted}}{\text{Diagnosis dates annotated in the dataset}} \qquad (2)$$

$$\textbf{F-score} = 2 * \frac{\text{Precision*Recall}}{\text{Precision + Recall}} \qquad (3)$$

C. Results

Table 3 shows the results obtained in the task of diagnosis date extraction. According to Table 3, a 84% F-score was obtained, in Test 4, while in Test 1 a 54% F-score was obtained. This indicates the impact of using the disambiguation process to extract the diagnosis date.

Moreover, Tests 2, Test 3, and Test 4 obtained better results than those obtained from Test 1. This suggests that using only the NLP-based annotators is not sufficient to extract the diagnosis date correctly.

According to Table 3, adding the Negation disambiguation component improves 5% in F-score, while adding the Speculation component improves 21% in F-score. This suggests that the diagnosis date extraction is more sensitive to speculation than to negation.

TABLE III
RESULTS OBTAINED TO DATE EXTRACTION

	Test 1	Test 2	Test 3	Test 4
Precision	0.56	0.62	0.82	**0.87**
Recall	0.52	0.55	0.79	**0.82**
F score	0.54	0.59	0.80	**0.84**

Finally, Figure 4 shows a F-score curve for all performed tests. These curves describe the impact of adding the Disambiguation components in the process to extract the diagnosis date.

Although all components improve diagnosis date extraction in comparison to using only just NLP-based annotators, it can be seen that the component with the highest rate of improvement is the Speculation component. Figure 4 shows that the most significant variation at the end of the curve occurs when Test 3 is performed. This fact indicates that filtering speculative annotations is a key component for extracting the diagnosis date correctly. Also, when negation annotations and family history annotations are filtered, similar improvement rates are obtained.

D. Error Analysis

We found the following limitations in our approach:

- Currently, lung cancer and dates are annotated separately. However, there are some sentences where it is also needed to recognize the relationships between these annotated concepts to improve extraction precision.
- Speculation is a linguistic phenomenon that should be studied more in detail in clinical notes written in Spanish. Some errors for extracting the diagnosis date are due to not detecting all the ways in which physicians express speculative findings.

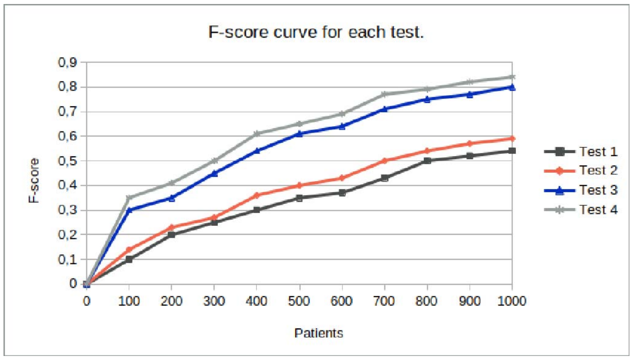

Fig. 4. F-score curve for all performed Tests

- We could not find a baseline proposal to compare our study in the diagnosis date extraction task for Spanish.

VI. Conclusions and future work

In this paper, an approach to find the lung cancer diagnosis and diagnosis date from multiple clinical notes written in Spanish has been proposed. The approach enhances diagnosis extraction using a disambiguation process and a rule-based algorithm. Knowing the exact diagnose date is paramount to be able to calculate patient survivorship.

Performed tests showed that the NLP-based annotation process can be improved using a post-processing phase in order to disambiguate annotations. Different tests have been conducted to analyze the contribution of different components such as negation, speculation, and the subject of the sentence to disambiguate diagnosis annotations. Results show the importance of these components to improve lung cancer diagnosis extraction.

Automatically extracting useful information from clinical notes, and in particular, accurate cancer diagnosis-related information, is a promising task to improve clinical decision support systems. The ability to analyze clinical texts written in Spanish opens great opportunities to develop more NLP-based clinical applications. In future works, we will explore other applications of the proposed approach.

Acknowledgment

This paper is supported by European Union's Horizon 2020 research and innovation program under grant agreement No. 727658, project IASIS (Integration and analysis of heterogeneous big data for precision medicine and suggested treatments for different types of patients).

References

[1] P. M. de Groot, C. C. Wu, B. W. Carter, and R. F. Munden, "The epidemiology of lung cancer," *Translational Lung Cancer Research*, vol. 7, no. 3, 2018. [Online]. Available: http://tlcr.amegroups.com/article/view/21996

[2] "Lung Health and Diseases lung disease lookup," https://www.lung.org/lung-health-and-diseases/lung-disease-lookup/lung-cancer/resource-library/lung-cancer-fact-sheet.html, accessed: 2020-01-30.

[3] "Lung Health and Diseases lung disease lookup," https://www.cancer.org/cancer/lung-cancer/about/key-statistics.html, accessed: 2020-02-14.

[4] L. Wang, L. Luo, Y. Wang, J. Wampfler, P. Yang, and H. Liu, "Natural language processing for populating lung cancer clinical research data," *BMC Medical Informatics and Decision Making*, vol. 19, no. Suppl 5, pp. 1–10, 2019. [Online]. Available: http://dx.doi.org/10.1186/s12911-019-0931-8

[5] M. Najafabadipour, J. M. Tuñas, A. Rodríguez-González, and E. Menasalvas, "Lung cancer concept annotation from spanish clinical narratives," in *Data Integration in the Life Sciences*, S. Auer and M.-E. Vidal, Eds. Springer International Publishing, 2019, pp. 153–163.

[6] I. Spasić, J. Livsey, J. A. Keane, and G. Nenadić, "Text mining of cancer-related information: Review of current status and future directions," *International Journal of Medical Informatics*, vol. 83, no. 9, pp. 605 – 623, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1386505614001105

[7] D. Demner-Fushman, W. W. Chapman, and C. J. McDonald, "What can natural language processing do for clinical decision support?" *Journal of Biomedical Informatics*, vol. 42, no. 5, pp. 760 – 772, 2009, biomedical Natural Language Processing. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1532046409001087

[8] Z. Q.T., S. Goryachev, and S. Weiss, "Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system," *BMC Medical Informatics and Decision Making*, vol. 30, no. 6, pp. 327–348, JUl 2006.

[9] "Snomed kernel description," http://http://www.snomed.org, accessed: 2020-01-04.

[10] D. Ferrucci and A. Lally, "UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment," *Natural Language Engineering*, vol. 10, no. 3-4, pp. 327–348, sep 2004.

[11] W. W. Yim, M. Yetisgen, W. P. Harris, and W. K. Sharon, "Natural Language Processing in Oncology Review," *JAMA Oncology*, vol. 2, no. 6, pp. 797–804, 2016.

[12] J. L. Warner, M. A. Levy, M. N. Neuss, J. L. Warner, M. A. Levy, and M. N. Neuss, "ReCAP: Feasibility and Accuracy of Extracting Cancer Stage Information From Narrative Electronic Health Record Data," *Journal of Oncology Practice*, vol. 12, no. 2, pp. 157–158, 2016.

[13] P. R. Deshmukh and R. Phalnikar, "Tnm cancer stage detection from unstructured pathology reports of breast cancer patients," in *Proceeding of International Conference on Computational Science and Applications*, S. Bhalla, P. Kwan, M. Bedekar, R. Phalnikar, and S. Sirsikar, Eds. Singapore: Springer Singapore, 2020, pp. 411–418.

[14] S. Ananiadou and J. C. Park, *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science): Introduction*. Springer, 2005, vol. 3248.

[15] T. L. Evans, P. E. Gabriel, and L. N. Shulman, "Cancer Staging in Electronic Health Records: Strategies to Improve Documentation of These Critical Data," *Journal of Oncology Practice*, vol. 12, no. 2, pp. 137–139, 2016.

[16] O. Bodenreider, "The unified medical language system: integrating biomedical terminology," *PubMed Central*, vol. 32, no. 3-4, pp. 327–348, Jan 2004.

[17] Z. Zeng, S. Espino, A. Roy, X. Li, S. A. Khan, S. E. Clare, X. Jiang, R. Neapolitan, and Y. Luo, "Using natural language processing and machine learning to identify breast cancer local recurrence," *BMC Bioinformatics*, vol. 19, no. Suppl 17, 2018.

[18] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute, "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 507–513, 2010.

[19] A. R. Aronson and F. M. Lang, "An overview of MetaMap: Historical perspective and recent advances," *Journal of the American Medical Informatics Association*, vol. 17, no. 3, pp. 229–236, 2010.

[20] A. Névéol, H. Dalianis, S. Velupillai, G. Savova, and P. Zweigenbaum, "Clinical Natural Language Processing in languages other than English: Opportunities and challenges," *Journal of Biomedical Semantics*, vol. 9, no. 1, pp. 1–13, 2018.

[21] M. Najafabadipour, M. Zanin, A. Rodriguez-Gonzalez, C. Gonzalo-Martin, B. N. Garcia, V. Calvo, J. L. C. Bermudez, M. Provencio, and E. Menasalvas, "Recognition of time expressions in Spanish electronic health records," *Proceedings - IEEE Symposium on Computer-Based Medical Systems*, vol. 2019-June, pp. 69–74, 2019.

[22] S. Sheikhalishahi, R. Miotto, J. T. Dudley, A. Lavelli, F. Rinaldi, and V. Osmani, "Natural language processing of clinical notes on chronic diseases: Systematic review," *Journal of Medical Internet Research*, vol. 21, no. 5, pp. 1–18, 2019.

[23] T. J. Mensalvas E and B. Guzman, "Profiling Lung Cancer Patients Using Electronic Health Records Environment," *Journal of Medical Systems*, vol. 42, no. 7, pp. 327–348, sep 2018.

[24] M. Najafabadipour, M. Zanin, A. Rodríguez-González, C. Gonzalo-Martín, B. Nuñez García, V. Calvo, J. Luis Cruz Bermudez, M. Provencio, and E. Menasalvas, "Recognition of time expressions in spanish electronic health records," in *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, June 2019, pp. 69–74.

[25] O. Solarte-Pabón, E. Menasalvas, and A. Rodriguez-González, "Spa-neg: an approach for negation detection in clinical text written in spanish," in *Proceeding Bioinformatics and Biomedical Engineering. IWBBIO*. Springer, Jan. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3625499